Variable length testing using the ordinal regression model.

Smits, N.; Finkelman, M.D.

published in
Statistics in Medicine
2014

DOI (link to publisher)
10.1002/sim.5936

document version
Publisher's PDF, also known as Version of record

citation for published version (APA)
Smits, N., & Finkelman, M. D. (2014). Variable length testing using the ordinal regression model. *Statistics in Medicine*, *33*(3), 488-499. https://doi.org/10.1002/sim.5936

Research Article

Statistics in Medicine

Received 14 November 2012, Accepted 11 July 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5936

Variable length testing using the ordinal regression model

Niels Smits[a*†] **and Matthew D. Finkelman**[b]

Health questionnaires are often built up from sets of questions that are totaled to obtain a sum score. An important consideration in designing questionnaires is to minimize respondent burden. An increasingly popular method for efficient measurement is computerized adaptive testing; unfortunately, many health questionnaires do not meet the requirements for this method. In this paper, a new sequential method for efficiently obtaining sum scores via the computer is introduced, which does not have such requirements and is based on the ordinal regression model. In the assessment, future scores are predicted from past responses, and when an acceptable level of uncertainty is achieved, the procedure is terminated. Two simulation studies were performed to illustrate the usefulness of the procedure. The first used artificially generated symptom scores, and the second was a post hoc simulation using real responses on the Center for Epidemiologic Studies Depression scale. In both studies, the sequential method substantially reduced the respondent burden while maintaining a high sum score quality. Benefits and limitations of this new methodology are discussed.

Keywords: computerized testing; ordinal regression; respondent burden; interim analysis

1. Introduction

Measurement plays a central role in both medical practice and research. It provides the basis of diagnosis and prognosis of patients, and evaluation of medical treatments [1]. Questionnaires have become an important method for measuring health-related constructs. In the past two decades, the number of bio-medical publications citing the word *questionnaire* has risen exponentially [2]; for the year 2012, PubMed© lists more than 38,000 citations. Questionnaires are used in a variety of medical fields. For example, Grotle *et al.* [3] developed a questionnaire for screening for chronic cases among lower back pain patients. Williams *et al.* [4] assessed the performance of a questionnaire to assess the impact of menopausal symptoms on health-related quality of life in a large US population-based study. Schulze *et al.* [5] used questionnaire data to assess major risk factors for the screening of individuals at high risk of developing type 2 diabetes in the general population. Often, questionnaires are built up from a set of questions, which are to be evaluated on a rating scale, and the resulting scores are added up to obtain a (possibly weighted) sum score. For example, the items of the Center for Epidemiologic Studies Depression (CES-D) [6] scale are totaled and the sum score is used as a measure of the severity of depression.

An important point in designing a health questionnaire is its length [7]. Although longer questionnaires may provide more data per patient and more reliable (i.e. internally consistent) scales, they also increase *respondent burden*. A greater respondent burden may not only reduce the quality of responses provided [8], but may also reduce the willingness to fill out the questionnaire at all [9, 10]. Therefore, for a questionnaire to be useful, it should not be unnecessarily time-consuming. A method for dealing with this issue is computerized adaptive testing (CAT), which has recently become very popular in the medical field [11–14]. CAT involves the administration of a questionnaire via the computer, and rather than using a fixed number of items, measurement uncertainty is fixed to an acceptable level, allowing

[a] *VU University Amsterdam, The Netherlands*
[b] *Tufts University School of Dental Medicine, Boston, U.S.A.*
**Correspondence to: Niels Smits, Department of Clinical Psychology, Faculty of Psychology and Education, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands.*
[†] *E-mail: n.smits@vu.nl*

the number of items administered to vary among respondents [15]; typically, CATs use substantially less items than the corresponding regular questionnaire [16]. Unfortunately, not every health questionnaire is suitable for a conversion toward CAT because a measurement model, which lies at the core of CAT, may not be an appropriate representation of the relationship between the construct of interest and the items. In measurement models, the covariance among items is explained by appointing the construct as a common cause of the item scores. Sometimes, questionnaire items do not meet the requirements of the measurement model because instead of a single anticipated construct, they measure multiple constructs. For example, Gao *et al.* [17] showed that the items of the General Health Questionnaire had a multidimensional structure, which would disqualify it as an input for a unidimensional CAT version. In such cases, to deal with the issue, researchers may make an effort to implement multidimensional CAT [18–20], although this may require a substantial investment. A more fundamental problem is that instead of being indicators of it, questionnaire items may *cause* the construct [21–24]. For example, quality of life is the common effect (i.e., an index score) of observable symptoms such as 'pain', and a poor quality of life does, therefore, not imply pain symptoms [25]. Likewise, instead of measuring a common construct, items may make up a causal chain [26, 27]. For example, in depression, one encounters symptoms such as 'lack of sleep', 'fatigue', and 'lack of concentration'; and instead of these symptoms being indicators of depression, it may be argued that lack of sleep leads to fatigue, which in turn causes lack of concentration [28].

In cases where questionnaire items are inappropriate for CAT, alternative methods for efficient assessment are needed. At first glance, such alternatives are readily available. For example: (i) short forms of the original questionnaires may be developed [17, 29]; (ii) incomplete designs may be used in which respondents only fill out a subset of the items [10, 30, 31]; or (iii) a decision tree may be used in which respondents only fill out items encountered in its branches [32]. These methods, however, do not have an equivalent to the conditional measurement precision of CAT. Consequently, it cannot be known if the uncertainty associated with a respondent's score meets some acceptable level.

Recently, two studies appeared introducing a new method for efficient health questionnaire administration, which did incorporate a measure of statistical uncertainty for the outcome. Finkelman *et al.* [33, 34] used stochastic curtailment, which is a method for computerized questionnaire administration designed for classification into categories, such as 'at risk' and 'not at risk'. Rather than using a measurement model, they specified a prediction model for forecasting observed class membership; the strategy is to stop testing when not yet administered items are unlikely to change the respondent's classification. During the assessment, after each item, the conditional probability that the full-length test will result in each classification ('at risk' and 'not at risk'), given the current set of responses, is estimated. It stops when the probability of either classification becomes higher than some threshold value representing an acceptable value of (un)certainty. Finkelman *et al.* performed real-data simulations, which showed that with this method, the number of items can be reduced substantially while maintaining a high classification accuracy.

In this paper, a method is presented that takes the idea of using probabilities to quantify outcome uncertainty from Finkelman *et al.* [33, 34], but which uses it for predicting sum scores rather than classifications. This paper has the following structure. We first begin with an introduction of the new approach. Then it is illustrated in two applications: one with artificial data and one with real data. Finally, benefits and limitations of this methodology are discussed.

2. Sequential testing using ordinal regression

Assume K Likert items with a number of categories M, scored on a 0 to $M-1$ rating scale. Let $(X_1, \ldots, X_K)$ be random variables representing the scores should all items be administered, and $(x_1, \ldots, x_K)$ their realized values. Let T denote the sum score for all K items, which varies between 0 and $K(M-1)$. During the assessment of a given respondent, $k < K$ items have been administered giving thus far a set of k item scores $\{x_i\}$, with $i = 1, \ldots, k$ indexing the order of administration of the items. This gives the cumulative score $c_k = \sum_{i=1}^{k} x_i$, and the remainder score $R_k = \sum_{i=k+1}^{K} X_i$, which is a random variable. The goal is to predict T on the basis of $\{x_i\}$. A first step is to specify a probability for each of the theoretical sum scores, given the scores thus far:

$$P(T = j \,|\{x_i\}), i = 1, 2, \ldots, k, j = 0, 1, \ldots, K(M-1) \tag{1}$$

Obviously, because c_k in $T = c_k + R_k$ is fixed, some sum scores j are impossible. For example, the highest theoretical sum score cannot be attained if at least one lower category score has been obtained thus far; likewise, the lowest theoretical sum score is impossible if at least one non-zero score has been obtained thus far. It is therefore more convenient to first consider probabilities for R_k:

$$P(R_k = j|\{x_i\}), i = 1, 2, \ldots, k, j = 0, 1, \ldots, (K-k)(M-1) \tag{2}$$

The resulting probabilities are then to be used as the probabilities for sum scores $j + c_k$. To obtain these probabilities, one can fit an appropriate statistical model in a large sample. For example, one can fit the proportional odds model (POM, [35, 36]), which is a cumulative logit model for ordinal responses:

$$P(R_k \leqslant j|\{x_i\}) = \left[1 + \exp\left(\alpha_j + \sum_{i=1}^{k} \beta_i x_i\right)\right]^{-1}, i = 1, 2, \ldots, k, j = 0, 1, \ldots, [(K-k)(M-1)-1] \tag{3}$$

where β_i is a parameter for item i, and α_j is a threshold for logit j. Note, that because the cumulative probability for the highest remainder score equals 1, the POM does not use it. Probabilities for each category are obtained by subtracting the cumulative probabilities, $P(R_k = j|\{x_i\}) = P(R_k \leqslant j|\{x_i\}) - P(R_k \leqslant j-1|\{x_i\})$. Given the current set of responses $\{x_i\}$, the estimated model provides probabilities $\hat{P}_j$. As a point prediction of R_k, one could use the category with the highest probability: $\arg\max_j \hat{P}_j$, and add this up to c_k to get a prediction of T. Alternatively, if one is willing to treat sums of item scores as *quantitative* rather than ordinal, using probabilities $\hat{P}_j$, a discrete density function of R_k can be constructed [37]. Using the expected value of this distribution as predicted value of the sum score gives:

$$\widehat{T_k} = c_k + \widehat{R_k} = c_k + \sum_j j\,\hat{P}_j \tag{4}$$

The shape of this predictive distribution is informative of the uncertainty associated with $\hat{T}$. A peaked distribution expresses that a few j are much more probable than others; a flat distribution expresses that all j are roughly equally probable. To quantify the variability of the predictive distribution, several measures may be used. For example, treating the remainder score as quantitative one could use the standard deviation of this distribution; treating it as nominal, one could use entropy [38]:

$$H_k = -\sum_j \hat{P}_j \log \hat{P}_j \tag{5}$$

High (low) values of H_k express high (low) variance, i.e., high (low) uncertainty for predicting T. In this paper, entropy is preferred over the standard deviation because of its interpretability as it quantifies the average missing information content when the value of a random variable is unknown. Like in CAT, uncertainty measure H can be fixed at an acceptable level, H^*, and item administration is continued until this level is reached, allowing for efficient assessment with adequate uncertainty for each respondent.

Applying the formulated method in a real assessment setting, would need two distinct construction phases. In the first phase, using a large sample from a relevant population, the ordinal regression parameters are estimated. In the second phase, the calibrated parameters are used to build the adaptive assessment. In the succeeding text, the two phases are described in more detail.

2.1. Calibration

Although some contiguous steps can be performed jointly, for clarity, they are discussed separately.

(1) Because it is important in a sequential method that the best items be administered first, the first step is to sort the K items from high to low in terms of their predictive power. There are several ways to do this. One could, for example, arrange the items according to the item-sum or item-remainder correlations. This approach, however, does not take into account that the set of items that forms the remainder score changes after each new item, and it may therefore be better to use forward selection. To that end, several sub-steps must be performed. In the first, K POMs

are estimated; the item for which the POM best predicts its remainder-score is selected. Next, $K-1$ POMs are estimated with for each model as dependent variable the remainder-score, which both omits the item in question, the item that was selected in the first step, and two predictors: the item in question and the item already selected. The item for which the POM best predicts the remainder-score is then selected. This sequence of model fitting is continued until the last item, but one is selected (evidently, the final item is not used as a predictor because there no longer is a remainder-score). Finally, arrange the items according to their rank in a data file, and if necessary, reverse-score negatively worded items.

(2) Start with the first item of the newly arranged data file as predictor of the remainder score. Subsequently, expand the model with one item at a time until $K-1$ models have been fitted. Obviously, in this sequence of models, each selected item changes from being an element in the remainder-score to being a predictor of the next remainder score. Store of each model both the regression parameters and the predictions for all observations in the calibration set. Note that the predicted *sum* score $\hat{T}_k$ can easily be obtained by adding cumulative score c_k to the predicted remainder score $\widehat{R_k}$. In addition, for the application to operate properly, estimates for all theoretically possible threshold parameters are needed. Consequently, if not all matching remainder scores are in the training data, artificial item scores, which yield these remainder scores should be added to the data set before fitting the model.

(3) Under each model, for all observations, store the expected value, $\hat{T}$ and entropy H.

(4) Study the bivariate distribution of true T and $\hat{T}$ as a function of H to choose an entropy of acceptable level, H^*.

2.2. *Application*

(1) Administer the items on a computer in the order of the arrangement described in section 2.1.

(2) After obtaining a score on item k, use $\{x_i\}$ and stored parameters of model k to compute H_k, and continue until it is lower than H^*.

(3) Compute $\hat{T}$ under the final model and store it (possibly after rounding it to integer).

A small example will illustrate the general idea. Suppose that a questionnaire contains five dichotomous items, each associated with a symptom, and that a large sample is used to calibrate the procedure; the ordering of items from high to low in terms of the predictive value is item 4, item 2, item 1, item 5, and item 3; and the parameters in Table I are the estimates of the four ($K-1=5-1=4$) POMs. Moreover, assume that the bivariate plots of T and $\hat{T}$ (e.g., resulting in a correlation of 0.95) would give a threshold for acceptable entropy $H^*=1.5$. A randomly chosen new respondent has scores 0,

Table I. The parameter estimates of four proportional odds models for a fictitious health questionnaire with five symptoms.

Model		α_j		β_i
$k=1$	0	−2.77	item 4	1.43
	1	−1.00		
	2	0.08		
	3	2.14		
$k=2$	0	−2.60	item 4	1.45
	1	−0.68	item 2	0.73
	2	0.89		
$k=3$	0	−1.40	item 4	1.14
	1	0.51	item 2	0.97
			item 1	0.54
$k=4$	0	0.18	item 4	1.15
			item 2	1.14
			item 1	0.37
			item 5	0.75

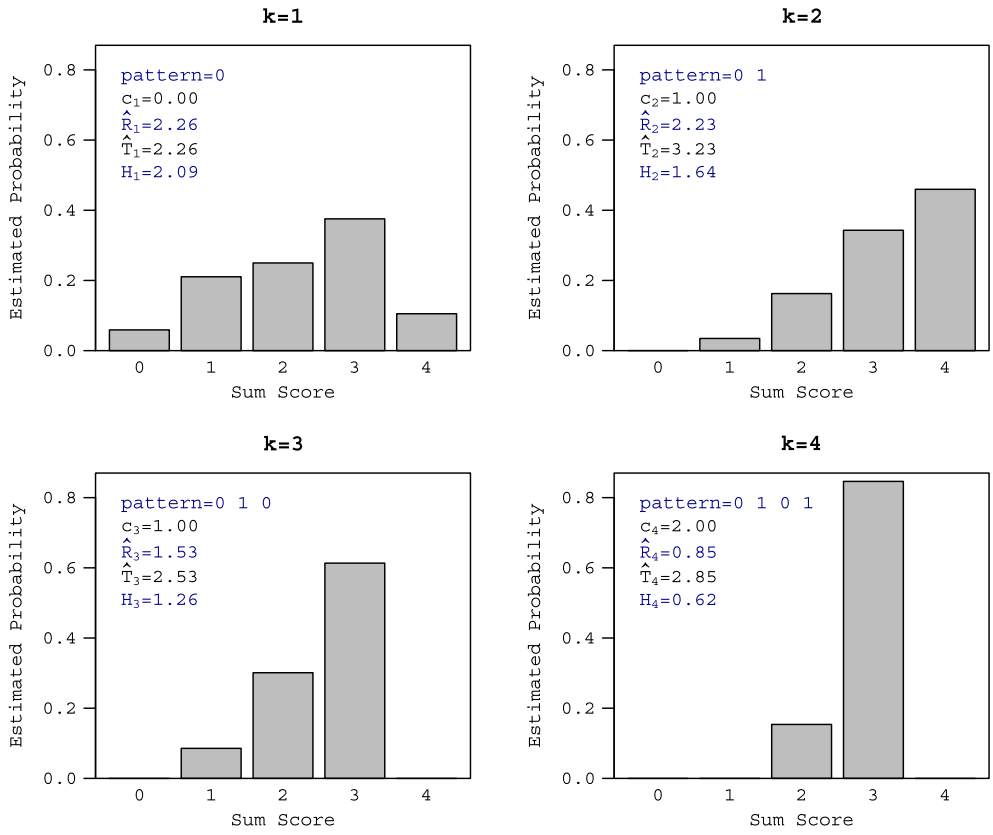

Figure 1. An illustration of the procedure for administering the items of a fictitious health questionnaire consisting of 5 symptom items for a theoretical respondent.

1, 1, 0, and 1 on items 1–5, respectively, should all items be administered. In the application, item 4 is administered first, yielding a score of 0. The parameters of the first model (see upper part of Table I) form the basis for the statistics in the upper left plot in Figure 1: the bars represent the estimated probabilities $\hat{P}_j$ for sum scores 0–4, $c_1 = 0$, $\widehat{R_1} = 2.26$, and therefore the predicted sum score $\widehat{T}_1 = 2.26$; the entropy of this predictive distribution is 2.09, which is larger than H^* by which a second item needs to be administered. Item 2 gives a score of 1, thus $c_2 = 1$, the parameters of the second model (see second part of Table I) give $\widehat{R_2} = 2.23$, and therefore $\widehat{T}_2 = 3.23$; the entropy of this predictive distribution is 1.64, which is larger still than H^*, and so a third item is administered. Next, the respondent scores a 0 on item 1, $c_3 = 1$, $\widehat{R_3} = 1.53$, $\widehat{T}_3 = 2.53$; the entropy is now 1.26, which is smaller than H^*, which stops the assessment. Instead of five items, three were administered, and the predicted sum score (2.53) is close to the actual sum score (3.00, in fact, after rounding to integer the two are equal). For the sake of completeness, the lower right panel of Figure 1 shows the predictions after the administration of four items, which would be necessary under a more stringent H^*. Note that if this criterion were still not met, all five items should be administered.

3. Simulation study using artificially generated item scores

To assess the usefulness of the new method, synthetic data sets were generated for which the measurement models forming the core of CAT would be inappropriate. Data files containing dichotomous symptom scores were generated using a structural model [24, 39]. In this model, symptoms were not manifestations of one or more factors, but instead, they were expressions of an underlying causal network [40, 41], and a sum score would represent a formative index [39]. Three sets consisting each of ten symptom items were created; the first set consisted of ten standard normally distributed exogenous variables, which had inter-item correlations of 0.25 on the population level. Next, to obtain scores on the second set, multivariate regression was used: the matrix with scores on the first set was post-multiplied by a 10×10 matrix of regression weights, and to the resulting predictions standard normally distributed

errors were added. The matrix of regression weights had the following structure: in each row and each column, five elements were fixed at zero, and each of the remaining five elements was set to a random draw from the standard uniform distribution. Consequently, each item in the second set was caused by five of the exogenous variables, and each exogenous variable caused five variables in the second set. Subsequently, in a similar fashion, the second set was used to obtain scores on the items of the third set. Next, all variables were converted to standard scores. The set of thirty simulated normally distributed scores represents the result of an unobservable causal process; to obtain observed responses for each item, a threshold was introduced, which divided the underlying continuum into scores that lead to two observed responses ('yes' and 'no') [42]. For each variable, a threshold value was drawn from a uniform $(-2, 2)$ distribution, and if the simulated score was above this threshold, it was recoded into a 1 ('yes'); if it was equal to or below the threshold, it was recoded into a 0 ('no').

For each data file, a full set of thirty symptom scores of one thousand simulees was generated. One hundred simulation replications were conducted to avoid imprecise results in estimation due to sampling error. The average sum score over the 100 data sets was 15.38 (SD = 1.98). This average, which quantifies the average number of reported symptoms (out of 30), shows that the current simulation population is similar to a clinical rather than a general population. The average within-data set inter-item correlation was 0.20 (over data files SD = 0.02), and the average Cronbach's α for the thirty-item scale was 0.89 (SD = 0.01). Note that, although internal consistency is high, the item scores result from a causal network, which makes them inappropriate as an input for CAT assessment.

In the calibration phase, in each data set, the scores of a randomly drawn five hundred simulees were used to estimate the parameters of the procedure. Because each data set had scores on 30 dichotomous symptoms, 29 POMs were estimated with 30 parameters in each model (the first model had 29 α_j's, and 1 β_i; the final model had 1 α_j and 29 β_i's). In the application phase, the remaining five hundred simulees were used as test set for the sequential procedure. No single value of H^* was selected for the application phase. Instead, to illustrate the impact of the stopping rule, the procedure was run under several entropy requirements (H^*s of 3.0, 2.5, 2.0, 1.5, 1.0, and 0.5). Table II shows the average outcomes under these six stopping rules. Note that all outcomes are average values over the 100 data files; e.g., the reported range is an average range, which explains the decimal values. Columns two to five provide statistics on respondent burden; evidently, the more stringent the stopping rule, the more items were administered. For example, under $H^* = 3.0$, on average, only about a fifth of the items was administered, whereas under $H^* = 1.0$ about 70% of the items were administered. Column six and seven give statistics on the relationship between T and $\hat{T}$; the Pearson correlation quantifies their linear dependency, and bias is their mean difference in a data set, which quantifies systematic under or over predictions. Obviously, for all stopping rules, the correlation is high, and bias is negligible, which is a prerequisite for the procedure to be useful. In addition, because in the medical field sum scores are not only used for assessing disease severity but also for screening purposes, classification outcomes are presented as well. Columns eight and nine provide statistics for concordance between full-length and early stopping classifications using an arbitrarily chosen cut score of 15 (which represents suffering from 50% of the symptoms). Classification overlap is the proportion of cases in which the two sum scores give identical classifications; κ is Cohen's [43] kappa coefficient for the agreement of nominal classifications of two sources. For all stopping rules, the classification overlap was at least 0.90, and kappa was at least 0.80, which according to rules of thumb specified by Landis and Koch [44] showed an 'almost perfect' agreement. To summarize

Table II. Average respondent burden and prediction results under six stopping rules over 100 synthetic multidimensional data files.

	Number of items used				Agreement observed and predicted sum score			
Stopping rule	Mean	SD	Median	Range	Correlation	Bias	Classification overlap[a]	κ^a
$H^* = 3.0$	5.95	2.51	5.05	3.86 – 13.11	0.94	−0.03	0.91	0.80
$H^* = 2.5$	10.61	3.52	9.40	7.21 – 20.76	0.96	−0.01	0.93	0.85
$H^* = 2.0$	15.18	3.65	13.86	11.73 – 24.93	0.98	−0.01	0.95	0.89
$H^* = 1.5$	18.56	3.64	17.31	14.75 – 27.45	0.99	0.01	0.97	0.94
$H^* = 1.0$	21.03	3.54	19.89	17.19 – 28.76	1.00	0.01	0.99	0.97
$H^* = 0.5$	23.50	2.99	22.72	19.88 – 29.91	1.00	0.00	0.99	0.99

[a] cut score is 15.

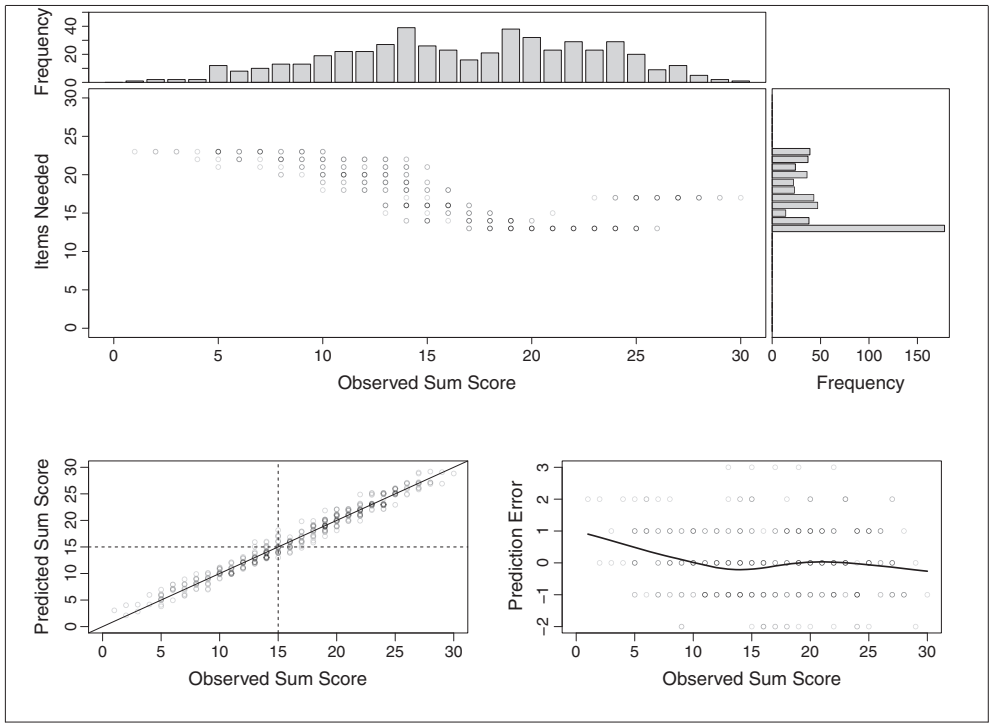

Figure 2. Outcomes under the '$H^* = 2.0$' stopping rule of one of the 100 simulated data sets. The upper panel depicts the number of items needed as a function of the observed sum score; bar plots of marginal distributions are also shown. The lower left panel displays the relationship between the observed and predicted sum scores (the solid line is the $y = x$ line; the dashed lines represent the cut score). The lower right panel shows the prediction error as a function of the observed sum score (the solid line is a Loess curve).

Table II briefly: for all stopping rules the accuracy of the predicted sum score was high, and with less stringent entropy requirements (administering fewer items), the obtained sum scores were somewhat more attenuated by prediction error, which resulted in somewhat lower correlations.

To shed more light on the procedure, several additional outcomes were inspected visually for each synthetic data file. Figure 2 shows the results of the test set under $H^* = 2.0$ of a randomly selected data file. The upper panel shows the number of items needed as a function of the true sum score and the marginal distribution of both variables. The sum score had a bell-shaped distribution, which shows that more extreme sum scores occurred less often. In addition, in this data set, respondents with scores closer to the mean needed fewer items than respondents with more extreme sum scores, although lower scores needed more items than higher scores. The lower left panel shows the scatter plot of observed and predicted sum scores, which shows a high correlation between the two. The lower right panel shows the prediction errors as a function of the observed sum score with a Loess curve added to it. As in most other data files, the curve deviates from a $y = 0$ line somewhat as both some overestimation for lower sum scores and some underestimation for higher sum scores occur, expressing a mild regression toward the mean in sum score prediction.

4. Post hoc simulation using actual Center for Epidemiologic Studies Depression item scores

In the second simulation study, data were taken from the Wisconsin Longitudinal Study [45], a study of the life course of a group of high school students who graduated in 1957. The present sample consisted of 6126 people aged between 64 and 65 years (53% women) who had complete data on all CES-D items. The CES-D scores were collected by telephone and mail interviews in 2003–2005. A more detailed description of the data set can be found on the accompanying website (http://www.ssc.wisc.edu/wlsresearch).

The CES-D [6] scale is commonly used both as an indicator of the severity of depression and as a first-stage screener for depression. It contains twenty items, and for each, the severity of a symptom is rated in terms of the number of days the respondent suffered from it in the past week. In the original article, Radloff [6] examined the factor structure of the scale in the general population and identified four factors, which would make it unsuitable for unidimensional CAT. In the standard version, symptom severity is expressed on a four-point rating scale with categories 0 days (0), 1−2 days (1), 3−4 days (2), and 5−7 days (3). Consequently its sum score ranges from 0 (no depressive complaints at all) to 60 (many depressive complaints). In the Wisconsin Longitudinal Study, an eight-point item scale was used: 0−7. These ratings were recoded to the original 0–3 scale. In the present sample, the sum score ranged from 0 to 53 (M=8.49, SD=7.81); the average inter-item correlation was 0.28, and Cronbach's α for the scale was 0.88. In addition, a principal components analysis on the polychoric correlation matrix showed that although a dominant factor explained most of the variance, multiple factors were needed to properly model the correlations, which suggests a multidimensional item set in the present population.

As in the first simulation study, the data set was randomly split into two parts (3063 observations each). The first sub-sample was used as a 'training set' to calibrate the parameters of the ordinal regression models. Because the CES-D has 20 items, 19 POMs were estimated, with a total of 760 parameters (the first model had 57 α_j's, and 1 β_i; the final model had 3 α_j's and 19 β_i's). The second set was used as a 'validation set' to study the performance of the proposed method in the application. Again, the procedure was run under several entropy requirements (H^*s of 3.5, 3.0, 2.5, 2.0, 1.5, and 1.0). The outcomes consisted of the same criterion variables as in the first study. Because screening for depression with the CES-D is commonly performed using a cut score of 16 [6], this value was selected for studying classification performance. Table III shows the outcomes under the six stopping rules. The pattern of outcomes was very similar to that of the first simulation study: under all stopping rules the accuracy of the predicted sum scores was high, and with higher H^* (administering fewer items), the obtained sum scores were slightly more attenuated by prediction error. For two stopping rules, $H^* = 2.0$ (0.20) and $H^* = 1.0$ (−0.18), the bias may seem somewhat further away from zero than for the other rules, but, fortunately, structural deviations of this size may be classified as negligible because they are quite small for scores ranging from 0 to 60. Figure 3 provides some more detail on the outcomes under the $H^* = 2.0$ stopping rule. The upper panel shows the number of items needed as a function of the true sum score and matching marginal distributions. Obviously, most respondents had low CES-D values: the modal value was 0, and scores were skewed toward the right. In addition, respondents with lower sum scores needed fewer items than respondents with higher scores; for example, sum scores of 0 needed 8 items, and sum scores $\geqslant 36$ needed 19 items. The lower left panel shows the scatter plot of observed and predicted sum scores, which shows a high linear dependency. The lower right panel shows the prediction errors as a function of the observed sum score with a Loess curve added to it. The plot shows that for lower scores, prediction errors are skewed toward the left (more under than over prediction), although the conditional means are close to 0. In addition, the Loess curve deviates from a $y = 0$ line somewhat, showing some overestimation for lower sum scores (especially scores 0–10), expressing a mild regression toward the mean in sum score prediction.

Table III. Respondent burden and prediction results for post hoc simulation on Center for Epidemiologic Studies Depression data (20 items).

	Number of items used				Agreement observed and predicted sum score			
Stopping rule	Mean	SD	Median	Range	Correlation	Bias	Classification overlap[a]	κ^a
$H^* = 3.5$	5.76	2.99	5	3 – 15	0.94	−0.01	0.96	0.83
$H^* = 3.0$	8.14	3.62	8	4 – 17	0.97	−0.07	0.97	0.87
$H^* = 2.5$	10.59	3.50	10	6 – 18	0.98	0.00	0.98	0.91
$H^* = 2.0$	12.65	3.09	13	8 – 19	0.99	0.20	0.98	0.94
$H^* = 1.5$	14.38	2.60	14	11 – 20	0.99	−0.09	0.99	0.96
$H^* = 1.0$	15.83	2.25	16	13 – 20	1.00	−0.18	1.00	0.99

[a] cut score is 16.

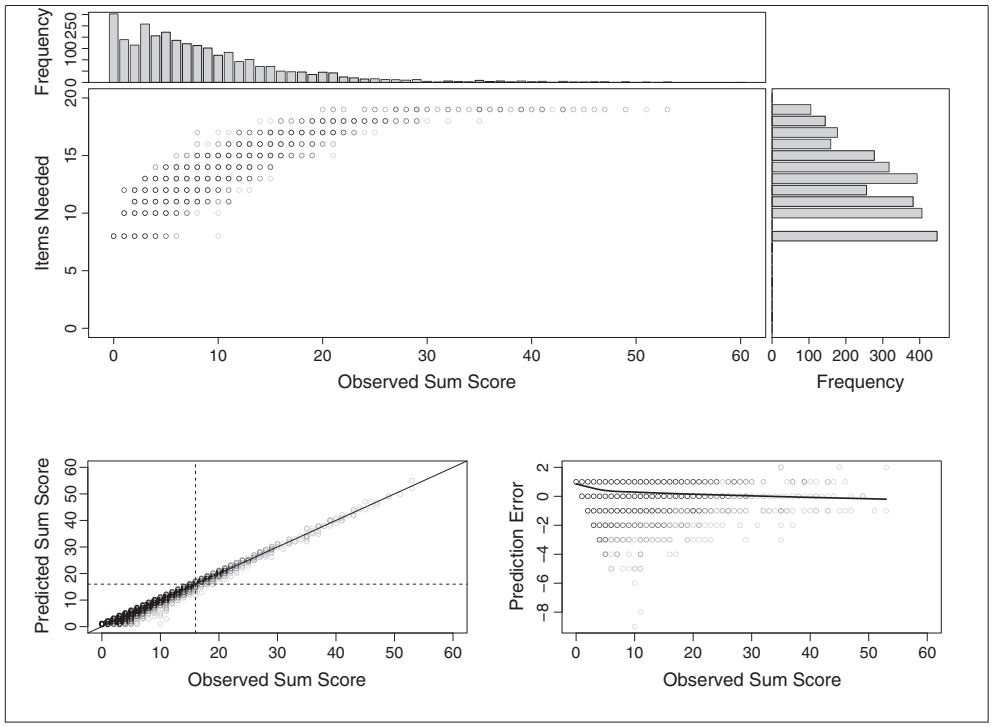

Figure 3. Outcomes under the '$H^* = 2.0$' stopping rule in the post hoc simulation using real Center for Epidemiologic Studies Depression scores. The upper panel depicts the number of items needed as a function of the observed sum score; bar plots of marginal distributions are also shown. The lower left panel displays the relationship between the observed and predicted sum scores (the solid line is the $y = x$ line; the dashed lines represent the cut score). The lower right panel shows the prediction error as a function of the observed sum score (the solid line is a Loess curve).

5. Discussion

In the present study a solution was sought for situations in which questionnaire items need efficient administration, but for which CAT is inappropriate. In contrast to CAT, the proposed variable length method has no specific requirements for the structure of the relatedness of questionnaire items; it only needs positive manifold. A series of prediction models was suggested as a vehicle to forecast full-length outcomes. Ordinal regression models are used to predict sum scores on the basis of already administered items, and the assessment is stopped when a pre-specified, acceptable, prediction uncertainty is met. In two simulation studies, under several levels of prediction uncertainty, predictions were evaluated at both sum score quality and classification accuracy. In both studies it was shown that when administering, on average, only a small number of items (e.g., 30% or 40%), sum score quality and classification accuracy remained high. In addition, the simulations suggested that the proposed method requires fewer items at the mode of the sum score distribution.

A further inspection of the outcomes suggested that the procedure somewhat suffers from scores being pulled toward the mean, a phenomenon which is also often encountered in CAT [46, 47]: lower values are somewhat overestimated, and higher values somewhat underestimated. The character and extent of this phenomenon appears to be associated with the distribution of sum scores in the population. In the first study, this distribution was bell-shaped and both overestimation and underestimation were found, whereas in the second study, it was skewed toward the right, and only underestimation was found. In the procedure, this bias may be reduced by using $\arg\max_j \hat{P}_j$ instead of an expected value for predicting sum scores, allowing for predictions, which are on either extreme of the scale.

Despite the presented outcomes, the reader may still be alarmed by the number of parameters that is needed for calibrating the procedure. It is true that this number is very high (e.g., 760 in the second simulation study), but it may be argued that the estimated models form a series of models based on forward selection, and therefore, the models are dependent; consequently, the *effective* number of parameters is much smaller. In addition, to study the effect of capitalization on chance in the calibration phase, we

compared the results in the validation set with those of the calibration set, and there was hardly a difference between the two (i.e., generalization error was negligible). For example, in the second simulation study, validation (Table III) and calibration results only differed in the third decimal for all outcomes. Naturally, there are ways to deal with the large number of parameters. For example, one could assume equidistant thresholds for the POM [35], or one could use unit weights for the predictors. Alternatively, one could use regularization methods, such as lasso, to reduce the number of predictors when estimating POMs. However, in model fitting, one is faced with a bias-variance tradeoff [48], and when the variance of the procedure is reduced by decreasing model complexity, its bias will in all likelihood increase, which is undesirable.

It should be noted that the new method is not presented as a replacement of CAT in cases where CAT would be appropriate. By contrast, it is especially useful when there is a need for efficient administration of a questionnaire that is unsuited for a conversion toward CAT because its items fail to comply with a measurement model. Moreover, the method may be used together with CAT, for example, when administering quality of life items. Quality of life questionnaires are often said to contain two types of items (e.g. [25, 49, 50]): causal indicators and effect indicators. Effect indicators, such as engaging in social activities, are manifestations of the level of quality of life; such items conform to a measurement model, which would allow for administering them using a unidimensional or multidimensional CAT. By contrast, causal indicators are assessments of symptoms of disease (e.g., loss of appetite), which affect the level of quality of life; these items do not comply with measurements models, and therefore, instead of CAT, the ordinal regression-based method would be used. In such a hybrid system, CAT would give measurements of the level of quality of life, and the new method would give sum scores for the symptoms of disease.

In the two presented simulation studies, we evaluated the method's performance from a statistical point of view; no attention was given to the coverage of health domains in the resulting assessments, however. In some instances, questionnaire designers may require that the content of items is balanced over the different domains [51], and that the assessment is not halted before items from each of the domains are administered. To control the number of items from each domain in the presented method, the content balancing rules for item ordering and early stopping, which are commonly applied in CAT (e.g., [52, 53]) can be easily adopted.

Note that the suggested method is population dependent, which means that the estimated parameters should only be applied to respondents for whom the calibration sample is representative. For example, if calibration is performed in a clinical population, it may be spurious to use it in a general population because the correlational structure of the items may be very dissimilar by which both sum score predictions and stopping rules may be inappropriate. If the method is to be used in a new population, it should be recalibrated to data from that same population.

This article is a first introduction of using ordinal regression models as a method for an efficient administration of health questionnaires. However, it suffers from some limitations. First, the variable length method was both applied upon synthetic data and applied post hoc on real data. Simulated and actual administrations could yield different results because respondents may behave differently in reality, although previous research shows that this difference is likely to be small [54]. Second, the programming and implementation of an actual early stopping procedure is more elaborate than a simulated procedure, especially if assessments should take place via the Internet. Third, as touched upon earlier, the method in its current form shows a mild regression toward the mean. Therefore, additional post hoc simulations are needed to study alternative methods for predicting sum scores under the POM. In addition, the procedure should be pilot tested using an actual sample to test the practicability of implementing it on the computer and to test the efficiency of actual assessments. Future research will address these and related issues.

Acknowledgements

This research uses data from the Wisconsin Longitudinal Study (WLS) of the University of Wisconsin-Madison. Since 1991, the WLS has been supported principally by the National Institute on Aging (AG-9775, AG-21079, and AG-033285), with additional support from the Vilas Estate Trust, the National Science Foundation, the Spencer Foundation, and the Graduate School of the University of Wisconsin-Madison. A public use file of data from the Wisconsin Longitudinal Study is available from the Wisconsin Longitudinal Study, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, Wisconsin 53706 and at http://www.ssc.wisc.edu/wlsresearch/data/. The opinions expressed herein are those of the authors.

References

1. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide*. Cambridge University Press: Cambridge, 2011.
2. Walter OB. Adaptive tests for measuring anxiety and depression. In *Elements of Adaptive Testing*, van der Linden WJ, Glas CAW (eds), Statistics for Social and Behavioral Sciences. Springer: New York, 2010; 123–136.
3. Grotle M, Vøllestad NK, Brox JI. Screening for yellow flags in first-time acute low back pain: reliability and validity of a norwegian version of the acute low back pain screening questionnaire. *The Clinical Journal of Pain* 2006; **22**(5): 458–467.
4. Williams RE, Levine KB, Kalilani L, Lewis J, Clark RV. Menopause-specific questionnaire assessment in US population-based study shows negative impact on health-related quality of life. *Maturitas* 2009; **62**(2):153–159.
5. Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, Möhlig M, Pfeiffer AFH, Spranger J, Thamer C, Häring HU. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes Care* 2007; **30**(3):510–515.
6. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement* 1977; **1**:385–401. DOI: 10.1177/014662167700100306.
7. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality of Life Research* 2002; **11**(3):193–205.
8. Herzog A, Bachman J. Effects of questionnaire length on response quality. *Public Opinion Quarterly* 1981; **45**(4):549–559.
9. Dillman DA, Sinclair MD, Clark JR. Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly* 1993; **57**:289–304.
10. Wacholder S, Carroll RJ, Pee D, Gail MH. The partial questionnaire design for case-control studies. *Statistics in Medicine* 1994; **13**:623–634.
11. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai J, Cella D. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care* 2007; **45**:S22–31.
12. Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: Time to let the CAT out of the bag? *Health Research and Educational Trust* 2005; **7**:347–364. DOI: 10.1111/j.1475-6773.2005.00446.x.
13. Walter OB, Becker J, Bjorner JB, Fliege H, Klapp BF, Rose M. Development and evaluation of a computer adaptive test for 'anxiety' (Anxiety-CAT). *Quality of Life Research* 2007; **16**:143–155. DOI: 10.1007/s11136-007-9191-7.
14. Jacobusse GJ, van Buuren S. Computerized adaptive testing for measuring development of young children. *Statistics in Medicine* 2007; **26**:2629–2638.
15. Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB, Kupfer DJ. Development of a computerized adaptive test for depression. *Archives of General Psychiatry* 2012; **69**(11):1104–1112.
16. Wainer H. *Computerized Adaptive Testing: A Primer*, 2nd ed. Lawrence Erlbaum Associates: Mahwah NJ, 2000.
17. Gao W, Stark D, Bennett MI, Siegert RJ, Murray S, Higginson IJ. Using the 12-item general health questionnaire to screen psychological distress from survivorship to end-of-life care: dimensionality and item quality. *Psycho-Oncology* 2011; **21**(9):954–961.
18. Gardner W, Kelleher KJ, Pajer KA. Multidimensional adaptive testing for mental health problems in primary care. *Medical Care* 2002; **40**(9):812–823.
19. Haley SM, Ni P, Ludlow LH, Fragala-Pinkham MA. Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the Pediatric Evaluation of Disability Inventory. *Archives of Physical Medicine and Rehabilitation* 2006; **87**(9):1223–1229.
20. Petersen MA, Groenvold M, Aaronson N, Fayers P, Sprangers M, Bjorner JB. Multidimensional computerized adaptive testing of the EORTC QLQ-C30: basic developments and evaluations. *Quality of Life Research* 2006; **15**(3):315–329.
21. Bollen K, Lennox R. Conventional wisdom on measurement: a structural equation perspective. *Psychological Bulletin* 1991; **110**:305–314.
22. Tien AY, Gallo JJ. Clinical diagnosis: a marker for disease? *The Journal of Nervous and Mental Disease* 1997; **185**(12):739–747.
23. Atkinson MJ, Lennox RD. Extending basic principles of measurement models to the design and validation of patient reported outcomes. *Health and Quality of Life Outcomes* 2006; **4**(1):65.
24. Edwards J, Bagozzi R. On the nature and direction of relationships between constructs and measures. *Psychological Methods* 2000; **5**(2):155–174.
25. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. *Quality of Life Research* 1997; **6**(2):139–150.
26. Borsboom D. Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology* 2008; **64**(9):1089–1108.
27. Borsboom D, Cramer AOJ, Schmittmann VD, Epskamp S, Waldorp LJ. The small world of psychopathology. *PloS ONE* 2011; **6**(11):e27407.
28. Schmittmann VD, Cramer AOJ, Waldorp LJ, Epskamp S, Kievit RA, Borsboom D. Deconstructing the construct: a network perspective on psychological phenomena. *New Ideas in Psychology* 2011; **30**:1–11.
29. Burisch M. Test length and validity revisited. *European Journal of Personality* 1997; **11**:303–315.
30. Raghunathan TE, Grizzle JE. A split questionnaire survey design. *Journal of the American Statistical Association* 1995; **90**:54–63.
31. Smits N, Cuijpers P, Beekman ATF, Smit JH. Reducing the length of mental health instruments through structurally incomplete designs. *International Journal of Methods in Psychiatric Research* 2007; **16**:150–160. DOI: 10.1002/mpr.223.
32. Yan D, Lewis C, Stocking M. Adaptive testing with regression trees in the presence of multidimensionality. *Journal of Educational and Behavioral Statistics* 2004; **29**:293–316. DOI: 10.3102/10769986029003293.

33. Finkelman MD, He Y, Kim W, Lai AM. Stochastic curtailment of health questionnaires: a method to reduce respondent burden. *Statistics in Medicine* 2011; **30**:1989–2004.
34. Finkelman MD, Smits N, Kim W, Riley B. Curtailment and stochastic curtailment to shorten the CES-D. *Applied Psychological Measurement* 2012; **36**:632–658.
35. Agresti A. *Categorical Data Analyses*, 2nd ed. Wiley: New York, 2002.
36. McCullagh P. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)* 1980; **42**:109–142.
37. Lacy M. An explained variation measure for ordinal response models with comparisons to other ordinal R^2 measures. *Sociological Methods & Research* 2006; **34**(4):469–520.
38. Gilula Z, Haberman SJ. Dispersion of categorical variables and penalty functions: derivation, estimation, and comparability. *Journal of the American Statistical Association* 1995; **90**(432):1447–1452.
39. Diamantopoulos A, Riefler P, Roth KP. Advancing formative measurement models. *Journal of Business Research* 2008; **61**(12):1203–1218.
40. Thagard P. *How Scientists Explain Disease*. Princeton University Press: Princeton, 2000.
41. McKay-Illari P, Russo F, Williamson J. *Causality in the Sciences*. Oxford University Press: Oxford, 2011.
42. Bock RD, Gibbons R, Muraki E. Full-information item factor analysis. *Applied Psychological Measurement* 1988; **12**(3):261–280.
43. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**:37–46.
44. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**:159–174.
45. Wisconsin Longitudinal Study. Wisconsin Longitudinal Study: graduates, siblings, and spouses. University of Wisconsin-Madison, Madison, WI, 2005. http://www.ssc.wisc.edu/wlsresearch/documentation.
46. Wang S, Wang T. Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement* 2001; **25**(4):317–331.
47. Embretson S, Reise SP. *Item Response Theory for Psychologists*. Lawrence Erlbaum: Mahwah, NJ, 2000.
48. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data mining, Inference and Prediction*. Springer: New York, 2001.
49. Boehmer S, Luszczynska A. Two kinds of items in quality of life instruments: 'indicator and causal variables' in the EORTC QLQ-C30. *Quality of Life Research* 2006; **15**(1):131–141.
50. Fayers PM, Hand DJ, Bjordal K, Groenvold M. Causal indicators in quality of life research. *Quality of Life Research* 1997; **6**(5):393–406.
51. Fayers PM. Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. *Quality of Life Research* 2007; **16**(1):187–194.
52. Van der Linden WJ. Optimal assembly of psychological and educational tests. *Applied Psychological Measurement* 1998; **22**(3):195–211.
53. Riley BB, Dennis ML, Conrad KJ. A comparison of content-balancing procedures for estimating multiple clinical domains in computerized adaptive testing: Relative precision, validity, and detection of persons with misfitting responses. *Applied Psychological Measurement* 2010; **34**(6):410–423.
54. Kocalevent RD, Rose M, Becker J, Walter OB, Fliege H, Bjorner JB, Kleiber D, Klapp BF. An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology* 2009; **62**:278–287. DOI: 10.1016/j.jclinepi.2008.03.003.